Les Cahiers du GERAD ISSN: 0711–2440

Global Stochastic Optimization with Low-Dispersion Point Sets

S. Yakowitz, P. L'Ecuyer
F. Vázquez-Abad

G–98–36

July 1998

Les textes publiés dans la série des rapports de recherche HEC n'engagent que la responsabilité de leurs auteurs. La publication de ces rapports de recherche bénéficie d'une subvention du Fonds F.C.A.R.

Global Stochastic Optimization with Low-Dispersion Point Sets

Sidney Yakowitz
Department of Systems and Industrial Engineering
University of Arizona, Tucson, AZ 85721, USA
yak@sie.arizona.edu

Pierre L'Ecuyer
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, C.P. 6128, succ. Centre-Ville
Montréal, CANADA, H3C 3J7
lecuyer@iro.umontreal.ca

Felisa Vázquez-Abad
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, C.P. 6128, succ. Centre-Ville
Montréal, CANADA, H3C 3J7
vazquez@iro.umontreal.ca

July, 1998

Les Cahiers du GERAD

G–98–36

Abstract

This study concerns a generic model-free stochastic optimization problem requiring the minimization of a risk function defined on a given bounded domain in a Euclidean space. Smoothness assumptions regarding the risk function are hypothesized and members of the underlying space of probabilities are presumed subject to a large deviation principle; however, the risk function may well be non-convex and multimodal. A general approach to finding the risk minimizer on the basis of decision/observation pairs is proposed. It consists of repeatedly observing pairs over a collection of design points. Principles are derived for choosing the number of these design points on the basis of an observation budget, and for allocating the observations between these points both in pre-scheduled and adaptive settings. On the basis of these principles, large-deviation type bounds of the minimizer in terms of sample size are established.

Résumé

Cette étude concerne un problème d'optimisation stochastique globale qui requiert la minimisation d'une fonction de risque définie sur un domaine borné de l'espace Euclidien. La fonction de risque est supposée suffisamment lisse et les variables aléatoires d'intérêt sont supposées suivre un principe des grandes déviations, mais la fonction de risque peut très bien être non convexe et multimodale. Nous proposons une approche générale pour minimizer la fonction de risque sur la base de paires décision/observation. La méthode consiste à observer une suite de paires à des points d'observation choisis dans l'espace de décisions. On établit des principes pour choisir le nombre de ces points d'observation, en fonction du budget de calcul disponible, et pour allouer les observations à ces points soit d'une manière pré-établie ou d'une manière adaptative. Sur la base de ces principes, on établit des bornes de type "grandes déviations" sur la distance entre la valeur de la fonction de risque au point retenu par l'algorithme et la valeur optimale de cette fonction.

Introduction

There are by now many areas of engineering and operations research in which optimization problems are precisely posed and completely modeled, but are difficult or impossible to resolve, either analytically or through conventional numerical analysis procedures. For example, the authors have encountered such situations in their studies of queue tuning, replacement strategies in reliability, intervention theory for epidemics, and optimization of machine learning codes for games and decisions. In these settings, models are available and one uses Monte Carlo (MC) simulation in conjunction with some sort of sequential optimization procedure. However, the procedures offered here are of the "machine learning" genre and thus have the additional potential of being applicable to actual experimental setups where one explores a response surface in the absence of a model.

Abstractly, the problems we have in mind involve minimizing a risk function over a given bounded real vector-valued domain. The stochastic optimization methods to solve this problem can be classified into two main categories:

(a) The methods based on *functional estimation*, which construct an estimate of the risk function over its entire domain, and then optimize the estimate.

(b) The *small-steps iterative* methods, which start at some initial design point, and change it by small amounts at successive iterations, on the basis of local information, such as a gradient estimator at each iteration.

We find in category (b) the *stochastic approximation* (SA) algorithm, with its several variants (e.g., Benveniste, Métivier, and Priouret 1990, Kushner and Clark 1978, Kushner and Vázquez-Abad 1996, Kushner and Yin 1997, L'Ecuyer and Yin 1998).

In the best situations, SA converges to the optimizer at the same rate, in terms of the total computing budget, as the MC method which estimates the risk at a single point. However, this is under several assumptions, such as unimodality and local convexity near the optimizer, that an "optimal" gain sequence is chosen, and that one has an "efficient" gradient estimator, e.g., unbiased with bounded variance (see L'Ecuyer and Yin 1998 for more details and slightly milder versions of these conditions). For certain classes of "smooth" systems, techniques such as perturbation analysis, score function or likelihood ratio methods, or finite differences with common random numbers and careful synchronization can provide a gradient estimator with the required efficiency (see Glasserman 1991, Glynn 1990, L'Ecuyer 1991, L'Ecuyer and Perron 1994, Rubinstein and Shapiro 1993). But these methods are often hard to implement and do not always apply. Then, one can often make do with straightforward finite differences and the less efficient Kiefer-Wolfowitz SA algorithm or its Spall (1992) variation. Another major difficulty with SA is the choice of the gain sequence. The algorithm performance is extremely sensitive to it and the optimal choice involves the Hessian of the risk function, which is typically unknown and hard to estimate. Finally, and perhaps more importantly, if the risk function has many local minima or flat areas, convergence may occur far away from the optimizer.

Approaches based on functional estimation (i.e., category (a)) are sometimes called *stochastic counterpart* methods, because they optimize a stochastic estimate of the risk function. Once the estimate has been constructed, it can be optimized by any deterministic optimization algorithm. For certain classes of problems where the risk depends on the probability law only, an estimator of the entire risk function can be obtained from a simulation at a single argument by using a likelihood ratio (Rubinstein and Shapiro 1993). This estimator may easily become unstable if the search domain is too large, as illustrated in, e.g., L'Ecuyer (1993) and Rubinstein and Shapiro (1993), so one would usually partition the search domain into smaller subsets, concentrate the sampling effort in the most promising areas, and so on. However, this likelihood ratio method is not universally applicable. A different functional estimation method for when a threshold parameter is sought, is proposed by L'Ecuyer and Vázquez-Abad (1997).

In principle, the techniques of nonparametric regression, which seek to approximate an unknown function over its entire domain solely on the basis of noisy observations at selected domain points are applicable to the stochastic counterpart approach (category (a), above). Müller (1985, 1989) explicitly follows the stochastic counterpart idea within the framework of the kernel regression technique. The nonparametric regression model is essentially that of stochastic optimization; there are a number of differences between the goals of his work and ours. In particular, Müller restricts attention to optimization over a finite *real interval*, and postulates moment conditions rather than large deviation principles. Whereas he too chooses a fixed-design viewpoint, only one observation per value is postulated. It is not easy to compare results because his convergence rates are couched in terms of the presumed order of the kernel function, and the criterion is different. Nevertheless, the rates of convergence are similar to ours in the non-adaptive setting. It is likely that a theory parallel to that of the present paper could be developed through the nonparametric regression principles, which have been nicely summarized in the text by Härdle (1989) and the earlier monograph by Prakasa Rao (1983).

The foundations of the present study include *nonparametric/non-Bayesian bandit theory*, stemming from Robbins (1952). It builds upon "off-line bandit" concepts in Yakowitz and Mai (1995) and a global stochastic approximation scheme in Yakowitz (1993). Lai and Yakowitz (1995) also investigate a finite-decision-space model using related methodology. The hypotheses there are weaker (no smoothness assumptions) and the convergence correspondingly slower. Dippon and Fabian (1994) give a globally-optimal algorithm combining nonparametric regression with stochastic approximation. Since the particular estimator is a partition-type rule, there is some overlap with the low-dispersion approach to follow. To our knowledge, the adaptive algorithm (Section 5) is new.

The present investigation similarly impinges on a line of study referred to as *ordinal optimization* by Ho and Larson (1995), and also the *ranking and selection* methods described in Chapter 9 of Law and Kelton (1991). These investigations fall into the bandit problem domain in that the goal is to choose the best, or alternatively a set of k decisions, one of which is best, on the basis of observed parameter/outcome pairs. In contrast to our developments, in many bandit investigations such as the preceding two references, the decision set is finite. Like our study, however, the discussion in Law and Kelton does give an assured quality of decision but under the hypothesis that the observations are Gaussian.

Further developments of the preceding line of inquiry are to be found in the literature of ranking and selection, within the framework of design and analysis of experiments (e.g. see the recent books by Bechhofer, Santner, and Goldsman 1995 or Mukhopadhyay and Solansky 1994). These studies differ from ours in that the decision space is finite and without any topological characteristics, the methodology is dependent on Gaussian theory, and the question of how many simulations to make is of major concern. Nevertheless, this literature could supplement or replace our techniques for selecting the best grid point. The aims of our adaptive approach (Section 5) do have some overlap with two-stage sampling procedures (e.g., Matejcik and Nelson 1995), which are also adaptive and intended to seek the point showing the most promise. On the other hand, the two-stage procedures have different motivations (namely, inferring the process variance and number of simulations needed for a given performance level).

The methodology to be related exploits the concept of *low-dispersion* point set used in quasirandom optimization to minimize the upper bound on discretization error (Niederreiter 1992 gives an overview and historical account of these ideas). The approach studied in this paper combines the ideas of *quasirandom search* for continuous deterministic global optimization and off-line *machine learning* for stochastic optimization over a finite set. These two techniques are explained in Niederreiter (1992) and Yakowitz and Mai (1995), respectively. The general outline is simple: Choose m points from the decision domain and perform r simulations to estimate the function at each of those points, then select the point with the least value of the function estimator. Questions of interest include:

- Assuming that N represents the total number of simulations that we are ready to perform, i.e., the computing budget, how should we choose m as a function of N?
- How should we select the m evaluation points within the decision domain?
- If the optimization error is defined as the difference between the risk at the selected point and the minimal risk over its entire domain, then at what rate does this error converge to zero with increasing computing budget, and under what conditions?
- Can the performance be improved by adaptive allocation of observations to the grid points?

We study those questions in this paper, under the following assumptions. Firstly, the probability law for the sample-mean of the risk estimator obeys an exponential bound as a function of sample size. This "moderate-deviation" assumption is satisfied by a collection of normal random variables with uniformly bounded variances, or any family of random variables with support on a bounded interval. As to be documented, it holds for a great many other random variables if the range of error is restricted to some sufficiently small interval. Secondly, we adopt a Lipschitz smoothness assumption near the optimizer.

We show that the risk of our decision strategy converges to zero in probability, and provide an assured rate for this convergence. Major features are that:

- The method requires no gradient estimator, only simulated realizations.
- The optimization is nonparametric: It does not require or use detailed information about the model structure and can therefore be used directly in an experimental or

on-line control setting. (Of course, if used in the simulation mode, a model must be specified in order to get the observations.)

- It is a global optimization method: It converges to the optimizer no matter how many local minima there are.
- The minimizer can be on the boundary as well as in the interior of the search domain.

This methodology is attractive because it is general and easy to implement. Moreover, it could be used to tune an actual system on-line without the need to perform modelling at all. It thereby constitutes a competitor to Kiefer-Wolfowitz SA and to other nonparametric machine learning methodologies.

1 The Optimization Problem

We want to solve the problem

$$\min_{\theta\in\Theta}\left(J(\theta) = \int_\Omega L(\theta,\omega)P_\theta(d\omega)\right), \tag{1}$$

where Θ is a compact region in the s-dimensional real space, $\{P_\theta,\ \theta\in\Theta\}$ is a family of probability measures over the sample space Ω, and L is a real-valued measurable function defined over $\Theta\times\Omega$. No closed-form expression is available for the function J. Suppose it can only be estimated by averaging i.i.d. replicates of $L(\theta) = L(\theta,\omega)$.

Let θ^* be an optimal solution to (1), i.e., a value of θ where the minimum is achieved, and let $J^* = J(\theta^*)$ be the optimal value. Given that we have a computing budget allowing N simulation runs, or allowing N learning observations if on-line, suppose we perform r runs at each point of the set $S_m = \{\theta_{m,1},\ldots,\theta_{m,m}\}\subset\Theta$, where $m = m_N$ is a non-decreasing function of N and (in terms for the "floor" function) $r = \lfloor N/m\rfloor$. Let

$$J_N^* = \min_{1\le i\le m} J(\theta_{m,i})$$

be the optimal value of J over the set of evaluation points S_m, and θ_N^* a value of $\theta\in S_m$ where this minimum is attained. The difference $J_N^* - J^*$ is what we loose by minimizing J over S_m instead of over Θ. For each $\theta_{m,i}$ in S_m, let

$$\hat{J}(\theta_{m,i}) = \frac{1}{r}\sum_{j=1}^{r} L_{i,j}, \tag{2}$$

where $L_{i,1},\ldots,L_{i,r}$ are i.i.d. replicates of $L(\theta_{m,i})$ simulated under $P_{\theta_{m,i}}$. Let

$$\hat{\theta}_N^* = \arg\min_{\theta_{m,i}\in S_m} \hat{J}(\theta_{m,i}) \tag{3}$$

be the point of S_m with the best empirical performance (in case of ties, choose any of the co-leaders). The point $\hat{\theta}_N^*$ is the one selected by the algorithm to "approximate" θ^*. What matters to us is not the distance between θ^* and $\hat{\theta}_N^*$, but the difference between the values of J at those points. Thus the performance of the algorithm is measured by the *error*:

$$\Delta_N = J(\hat{\theta}_N^*) - J^*. \tag{4}$$

This error is a random variable and will be bounded only in a probabilistic sense. We are interested in its convergence rate to zero.

The error is a sum of two components: the *discretization error* $J_N^* - J^*$ and the *selection error* $J(\hat{\theta}_N^*) - J_N^*$. The latter is (stochastically) reduced by increasing r and the former by increasing m. So for a given N, there is a tradeoff to be made. If m is large and r small, the discretization error is small, but one is likely to select a "bad" point in S_m, because of large errors in the estimators $\hat{J}(\theta_{m,i})$. Alternatively, if m is small and r is large, the chances are good that $\hat{\theta}_N^*$ is the best value among the points of S_m, i.e., that $J(\hat{\theta}_N^*) = J_N^*$, but since S_m is too sparse, the optimal value J^* might be quite a bit lower than J_N^*. Theorem 1 (in Section 3) will give us a good tradeoff by increasing m at a rate just high enough so that the probability that the selection error exceeds the discretization error diminishes to 0.

2 Low-Dispersion Point Sets

We now examine how to choose S_m. Let $\|\cdot\|_p$ be the L_p norm on $\mathbb{R}^s$, for $1 \le p \le \infty$. For example, $p = 2$ gives the Euclidean norm and $p = \infty$ gives the sup norm defined by $\|(x_1, \ldots, x_s)\|_\infty = \max(|x_1|, \ldots, |x_s|)$. We assume that Θ is compact in $\mathbb{R}^s$. Let $B_p(\theta, t) = \{x \in \mathbb{R}^s \mid \|x - \theta\|_p \le t\}$, the closed ball of radius t centered at θ. The *dispersion* (or *covering radius*) of a set of points $S_m = \{\theta_{m,1}, \ldots, \theta_{m,m}\}$ in Θ, with respect to the L_p norm, is defined as

$$d_p(S_m, \Theta) = \sup_{\theta \in \Theta} \min_{1 \le i \le m} \|\theta - \theta_{m,i}\|_p. \tag{5}$$

It is the minimal value of t such that the balls $B_p(\theta_{m,1}, t), \ldots, B_p(\theta_{m,m}, t)$ cover Θ. Define

$$H_p(t) = \sup_{\theta \in B_p(\theta^*, t) \cap \Theta} (J(\theta) - J(\theta^*)). \tag{6}$$

Proposition 1 *For $1 \le p \le \infty$, the discretization error is bounded by*

$$J_N^* - J^* \le H_p(d_p(S_m, \Theta)). \tag{7}$$

Proof. Since $\theta^* \in \Theta$, there is at least one point $\theta_{m,i_0} \in S_m$ such that $\|\theta^* - \theta_{m,i_0}\|_p \le d_p(S_m, \Theta)$. By the definition of H_p, one has $J(\theta_{m,i_0}) - J(\theta^*) \le H_p(d_p(S_m, \Theta))$. But $J_N^* \le J(\theta_{m,i_0})$ and the conclusion follows. □

The upper bound (7) is tight in the sense that one can easily construct functions for which it is reached, for any given p. To minimize this bound, low-dispersion point sets are wanted. The bound also suggests that convergence should occur faster if $H_p(t)$ is small and flat near $t = 0$, because a small $H_p(t)$ means that $J(\theta)$ is close to $J(\theta^*)$ all over the ball $B_p(\theta^*, t)$, so the discretization error is small whenever S_m has a point in that ball. Note that $H_p(t)$ is increasing in p, whereas $d_p(S_m, \Theta)$ is decreasing in p, and there is no general rule telling which value of p gives the smallest upper bound. This is why we do not stick to a particular value of p in this paper.

Two simple choices for the point set S_m are: (1) a rectangular grid and (2) a random set of points. In the following examples, we look at what the dispersion is in these two cases, when Θ is the s-dimensional unit hypercube. We then look at the discretization error when the function is locally quadratic near the optimizer (a common assumption in optimization).

Example 1 Let $d_\infty(S_m) = d_\infty(S_m, [0,1]^s)$ denote the dispersion of S_m over the s-dimensional unit hypercube, with the sup norm. Let $k = \lfloor m^{1/s} \rfloor$ and

$$S'_m = \left\{ (x_1, \dots, x_s) \mid x_j \in \left\{ \frac{1}{2k}, \frac{3}{2k}, \dots \frac{2k-1}{2k} \right\} \text{ for each } j \right\} \cup \Psi,$$

where Ψ is a set of $m - k^s$ arbitrary points in Θ. With this set, one has

$$d_\infty(S'_m) = \frac{1}{2k} = \frac{1}{2\lfloor m^{1/s} \rfloor}. \tag{8}$$

Sukharev (1971) shows that no other set S_m gives a lower value of $d_\infty(S_m)$ (see also Theorem 6.8 of Niederreiter 1992). The dispersion of S'_m with the L_p norm is

$$d_p(S'_m) = d_p(S'_m, [0,1]^s) = \frac{s^{1/p}}{2\lfloor m^{1/s} \rfloor}. \tag{9}$$

Example 2 Suppose that random points are generated independently and uniformly over $\Theta = [0,1]^s$, and let S_m be the set that contains the first m points. Then, with probability one, $d_\infty(S_m, \Theta) = O((\ln m/m)^{1/s})$. This result is proved by Deheuvels (1983). The implication is that for large m, random selection is slightly worse than rectangular grids or better pattern strategies. □

Example 3 Again let $\Theta = [0,1]^s$. Suppose that J has a quadratic upper bound, in the sense that $J(\theta) - J(\theta^*) \leq K(\|\theta - \theta^*\|_2)^2$ for some constant K. Then, $H_p(t) \leq \bar{H}_p(t) = Kt^2 s^{1-2/p}$ for $2 \leq p < \infty$ and $H_\infty \leq \bar{H}_\infty(t) = Kt^2 s$. If we use this with the point set S'_m, we obtain from (8) and (9) the upper bound on the discretization error:

$$\bar{H}_p(d_p(S'_m)) = \frac{Ks}{4\lfloor m^{1/s} \rfloor^2} \tag{10}$$

for $p \geq 2$. In this particular case, using any $p \geq 2$ in (7) gives the same bound. More generally, if $J(\theta) - J(\theta^*) \leq K(\|\theta - \theta^*\|_2)^q$ one gets the upper bound

$$\bar{H}_p(d_p(S'_m)) = K \left(\frac{\sqrt{s}}{2\lfloor m^{1/s} \rfloor} \right)^q. \tag{11}$$

for $p \geq 2$. □

In $s \geq 2$ dimensions, choosing S_m to minimize $d_2(S_m)$ over the unit hypercube is a hard unresolved problem and the optimal value of $d_2(S_m)$ is also unknown. Since $d_2(S_m) \geq d_\infty(S_m)$, (8) gives a lower bound on $d_2(S_m)$. For the s-dimensional *unit torus* $[0,1)^s$, the point sets with the lowest dispersion known, up to 22 dimensions, are the *Voronoi's principal lattices of the first type* (see Conway and Sloane 1988, p.115). For more about low-dispersion point sets, see also Niederreiter (1992) and the references given there.

3 Assumptions and Main Results

The next proposition provides a large deviation result for the selection error. We then build on this to obtain convergence rate results for the error Δ_N and study the question of how fast m should increase as a function of N. We need the following deviation assumption regarding replicated observations.

Assumption A1 (Noise 1) There are positive numbers R, ϵ_1, and κ, such that for all $r \geq R$, for all $0 < \epsilon \leq \epsilon_1$ and $\theta \in \Theta$,

$$P\left[\left|\frac{1}{r}\sum_{j=1}^{r} L_j - J(\theta)\right| > \epsilon\right] \leq e^{-r\kappa\epsilon^2}, \tag{12}$$

where $L_1, \ldots, L_r$ are r i.i.d. replicates of $L(\theta)$ under under P_θ,

This assumption obviously holds with $R = 1$ and κ not depending on ϵ_1 if $L(\theta)$ is normal with $\sup_{\theta\in\Theta} \mathrm{Var}[L(\theta)] < \infty$, and by the Bernstein inequality, if the supports of $L(\theta)$ are uniformly bounded in θ. It is well-known that other random variable families satisfy Assumption A1. It turns out to hold for many random variables $L(\theta)$ having a finite moment generating function, and, in fact, from Ellis (1985), p. 247, we have that if the moment generating function is finite for all real values, then for any particular θ, and $\sigma^2(\theta)$ the variance of L_1,

$$P_\theta\left[\left|\frac{1}{r}\sum_{j=1}^{r} L_j - J(\theta)\right| > \epsilon\right] \leq e^{-r\epsilon^2/(2\sigma^2(\theta))+O(r\epsilon^3)}, \tag{13}$$

so to assure validity of A1, it suffices to take κ a little smaller than one half the inverse of the largest (over Θ) variance of $L(\theta)$. From pp. 20 through 22 of Shwartz and Weiss (1995), one sees that the Poisson and exponential families also satisfy A1. Petrov (1975), p. 249, gives leading constants in the rate of convergence of (13) which are valid whenever the tails of the characteristic function of $L(\theta)$ are bounded away, in absolute value, from 1. Feller (1966), p. 520, also gives a related bound.

Proposition 2 *Under Assumption A1, with κ and ϵ_1 as defined there, one has that for some N_0 and all $N > N_0$,*

$$P\left[J(\hat\theta^*_N) - J^*_N > 2\epsilon\right] \leq me^{-r\kappa\epsilon^2}, \text{ for } 0 < \epsilon \leq \epsilon_1/2, \tag{14}$$

Proof. Note that if $|\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| \le \epsilon$ for $i = 1, \ldots, m$, then $J(\hat{\theta}^*_N) - J^*_N \le 2\epsilon$. Therefore, for N sufficiently large,

$$\begin{aligned} P\left[J(\hat{\theta}^*_N) - J^*_N > 2\epsilon\right] &\le P\left[|\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| > \epsilon \text{ for some } i\right] \\ &\le \sum_{i=1}^{m} P\left[|\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| > \epsilon\right] \\ &\le me^{-r\kappa\epsilon^2}, \end{aligned}$$

where the last inequality follows from Assumption A1. □

Corollary 1 *Under Assumption A1 and its notation, one has that for some N_0, for all $N > N_0$ and $0 \le \epsilon \le \epsilon_1/2$,*

$$P\left[\Delta_N > 2\epsilon + H_p(d_p(S_m, \Theta))\right] \le me^{-r\kappa\epsilon^2}, \tag{15}$$

for all $p \ge 1$, where H_p is defined in (6).

Proof. Write the error Δ_N as the sum of its two (non-negative) components $J^*_N - J^*$ and $J(\hat{\theta}^*_N) - J^*_N$. The probabilities of these two components are bounded by (7) and (14), respectively, which yields the result. □

The quality of the bound (15) as a function of N depends on the behavior of $H_p(t)$ as a function of t, of $d_p(S_m, \Theta)$ as a function of m, and of m as a function of N. To proceed further, we thus need assumptions about the rate of increase of J around the optimizer and about the dispersion. Choose any p, $1 \le p \le \infty$.

Assumption A2 (Smoothness) One has $H_p(t) \le K_1 t^q$ for $t \le t_0$, for some positive constants K_1, q, and t_0.

Assumption A3 (Dispersion) One has $d_p(S_m, \Theta) \le K_2/\lfloor m^{1/s} \rfloor$, where K_2 is a constant.

The constants K_1 and K_2 in the assumptions may depend on s, but not on m. Assumption A2 holds in particular for $q = 1$ if J has a bounded gradient near θ^* and for $q = 2$ if it is locally quadratic. If Θ is the unit hypercube $[0,1]^s$, Example 1 shows how to select S_m so that A3 holds with $K_2 = 1/2$ for $p = \infty$ and with $K_2 = s^{1/p}/2$ for $1 \le p < \infty$.

For each positive integer N and each constant $C > 0$, let $m^*_N(C)$ be the largest integer m such that $m \le C \cdot (N/\ln N)^{s/(s+2q)}$ and $\lfloor N/m \rfloor \ge C^{-1-2q/s} \lfloor m^{1/s} \rfloor^{2q} \ln N$. Observe that

$$m^*_N(C) \sim C \cdot \left(\frac{N}{\ln N}\right)^{\frac{s}{s+2q}} \tag{16}$$

and

$$\lfloor (m^*_N(C))^{1/s} \rfloor^{-q} \sim C^{-q/s} \left(\frac{N}{\ln N}\right)^{\frac{-q}{s+2q}}.$$

As indicated by the next theorem, the latter expression is an upper bound on the rate of convergence in probability of the error Δ_N, when $m = m^*_N(C)$. This bound improves as q increases and deteriorates as s increases.

Theorem 1 *Let Conditions A1–A3 be in force for a given p and suppose that $m = m_N^*(C)$. Then, there are two constants K_0 and N_0 (which may depend on s and q) such that for all $N \geq N_0$,*

$$P\left[\Delta_N > K_0(N/\ln N)^{-q/(s+2q)}\right] \leq C(\ln N)^{-s/(s+2q)}. \tag{17}$$

Proof. Presume N_0 is sufficiently large that, for ϵ_1 as in Assumption A1 and $N \geq N_0$,

$$K_0(N/\ln N)^{-q/(s+2q)} < \epsilon_1/2$$

regardless of what K_0, to be specified later, turns out to be. Let K_3 be the constant satisfying $\kappa K_3^2 = s/(s+2q)$, and let $\epsilon = K_3\lfloor m^{1/s}\rfloor^{-q}C^{1/2+q/s}$. From A2 and A3, we have that $H_p(d_p(S_m, \Theta)) \leq K_1K_2^q\lfloor m^{1/s}\rfloor^{-q}$. So, this choice of ϵ makes the selection error decrease at the same rate as the discretization error, and therefore gives a good tradeoff in rates for the sum of these two errors. Let $K_4 = 2K_3C^{1/2+q/s} + K_1K_2^q$. Now, from (15) and by our choice of m,

$$\begin{aligned}
P\left[\Delta_N > K_4\lfloor m^{1/s}\rfloor^{-q}\right] &\leq P\left[\Delta_N > 2\epsilon + H_p(d_p(S_m, \Theta))\right] \\
&\leq me^{-\kappa\epsilon^2\lfloor N/m\rfloor} \\
&\leq me^{-\kappa K_3^2 \ln N} \\
&\leq C\cdot(N/\ln N)^{s/(s+2q)}e^{-(\ln N)s/(s+2q)} \\
&= C\cdot(\ln N)^{-s/(s+2q)}.
\end{aligned}$$

Since $\lfloor m^{1/s}\rfloor^{-q} \sim C^{-q/s}(N/\ln N)^{-q/(s+2q)}$, there are constants K_0 and subsequently N_0 such that $K_4\lfloor m^{1/s}\rfloor^{-q} \leq K_0(N/\ln N)^{-q/(s+2q)}$ for all $N \geq N_0$, and this completes the proof. □

Example 4 Suppose that the assumptions hold for $q = 2$. Then the convergence rate is $O_p((N/\ln N)^{-2/(s+4)})$. This gives $O_p((N/\ln N)^{-2/5})$ with $m \sim (N/\ln N)^{1/5}$ in one dimension, $O_p((N/\ln N)^{-1/3})$ with $m \sim (N/\ln N)^{1/3}$ in two dimensions, and so on. □

Example 5 Consider the random function

$$L(\theta) = (\theta_1 - 0.5)\ \sin(10\ \theta_1) + (\theta_2 + 0.5)\ \cos(5\theta_2) + N(0,1) \tag{18}$$

where $N(0,1)$ is a standard normal random variable, and $\theta = (\theta_1, \theta_2) \in [0,1]^2$. Suppose we want to minimize $J(\theta) = E[L(\theta)]$ over the set $[0,1]^2$. In dimension $s = 2$, this problem can easily be solved by finding the values of the objective function

$$J(\theta_1, \theta_2) = (\theta_1 - 0.5)\sin(10\ \theta_1) + (\theta_2 + 0.5)\cos(5\ \theta_2) \tag{19}$$

over a closely spaced uniform grid. By this means, we found that the optimal value of $J(\cdot)$ is approximately $J^* = -1.5020$. Figure 1 is a mesh diagram of this objective function.

In our computational study related to Theorem 1, we took

$$m_N^*(C) = \lfloor 10(N/\ln N)^{1/3}\rfloor,$$

and at each point, made $r = r(N) = \lfloor N/m_N^* \rfloor$ replications. The point set used was S'_m defined in Example 1, with m always equal to a square, namely $m = \lfloor\sqrt{m_N^*(C)}\rfloor^2$. After all this rounding, the nominal values of N = 100, 500, 1000, 5000, 10000, and 20000 turned out to require only the numbers $\tilde{N} = m(N)r(N)$ of observations shown in the leftmost column of Table 1. At each such value, we repeated the estimation of θ_N^* 10 times. That is, in 10 independent experiments, $\hat{\theta}_N^*$ was computed. The second column gives the number of these replications for which tolerance was exceeded, that is, for which $J(\hat{\theta}_N^*) - J_N^* > (N/\ln(N))^{-1/3}$. The next two columns give the average selection and discretization errors, respectively. Then comes the empirical standard deviation of $J(\hat{\theta}_N^*)$. The next column gives the numerical value of the tolerance. The final column is the number of points m. The average value of $J(\hat{\theta}_N^*) - J^*$ can be obtained easily by adding up columns 3 and 4. One can see from the table that the actual discretization error is not monotonically decreasing in m even if the upper bound in (7) is. After looking at the graph of the function in Figure 1, this is certainly not surprising. Even for a very small m, one can be lucky in having a point θ in S'_m for which $J(\theta)$ is very close to J^*. Nonetheless, for this example, selection and discretization errors are reasonably commensurate.

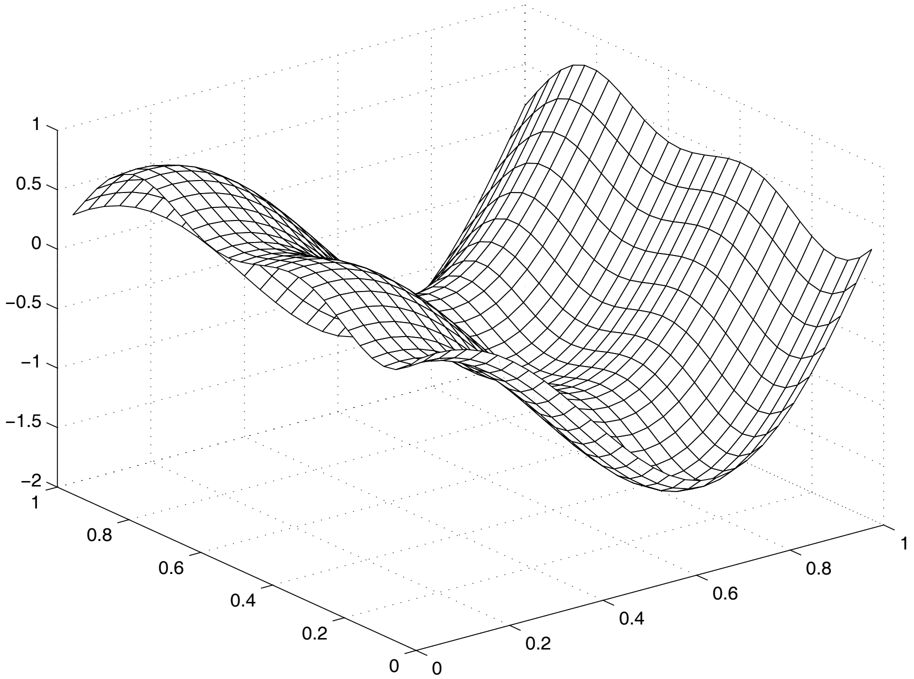

Figure 1: A mesh plot of the objective function $J(\theta)$

Theorem 1 does not give the fastest possible convergence. One could substitute in place of $\ln N$ in (17) a function growing more slowly in N, such as $\sqrt{\ln N}$. On the other hand, if one takes $m_N^*(C) \sim N^{s/(s+2q)}$ and $\epsilon = m^{-q/s}$, there is no assurance that convergence occurs; the probability bound $me^{-\kappa\epsilon^2\lfloor N/m \rfloor}$ grows with N. In short, there does not seem to be much ground for improvement in (17).

Our analysis does not give convergence rates of mean squared error performance under the noise assumption A1. On the other hand, we have noted in the case of Gaussian or bounded observations $L(\theta)$, that (12) holds for all $\epsilon > 0$ and all r.

Proposition 3 *Let Conditions A2 to A3 hold for some $p \geq 1$, and A1 hold for some fixed κ and for all $\epsilon > 0$. If $m_N^*(C) \sim C \cdot N^{s/(2s+2q)}$, then*

$$E[\Delta_N^2] = O(N^{-q/(q+s)}). \tag{20}$$

Proof. Since $E[\Delta_N^2] \leq E[(J(\hat{\theta}_N^*) - J_N^*)^2] + E[(J_N^* - J^*)^2]$, it suffices to bound each of these two expectations. As in the proof of Theorem 1, $H_p(d_p(S_m, \Theta)) \leq K\lfloor m^{1/s} \rfloor^{-q}$. Squaring this bound and using (7) shows that $E[(J_N^* - J^*)^2] = O(N^{-q/(q+s)})$. The selection error was shown in Proposition 2 to satisfy $P[|J(\hat{\theta}_N^*) - J_N^*| > 2\epsilon] \leq me^{-\kappa\epsilon^2\lfloor N/m \rfloor}$. Then, with $m = m_N^*(C)$,

$$\begin{aligned} E[(J(\hat{\theta}_N^*) - J_N^*)^2] &= \int_0^\infty P[(J(\hat{\theta}_N^*) - J_N^*)^2 > x]dx \\ &\leq \int_0^\infty me^{-\kappa\lfloor N/m \rfloor x/4}dx \\ &= \frac{4m}{\kappa\lfloor N/m \rfloor} \;=\; O(m^2/N) \;=\; O(N^{-q/(q+s)}). \end{aligned}$$

□

In the case that $s = 1$ and $q = 2$, this bound turns out to be $O(N^{-2/3})$.

4 Increasing N Dynamically and Low-Dispersion Sequences

In the setting of the previous sections, we were interested in how the error decreases with the computing budget N, assuming that N was fixed at the beginning of the experiment.

Suppose now that we do not fix the computing budget in advance, but reserve the right to stop at any value of N. For example, after N simulation runs, we may decide to go on for $N' - N$ additional runs. Of course, the data taken during the first N runs ought to be used effectively within the larger set of runs and this imposes strong restrictions on the choice of the point sets S_m as a function of N. More specifically, suppose that after N runs, the set of evaluation points is $S_m = \{\theta_1, \ldots, \theta_m\}$, with $r_{N,i}$ runs at point θ_i, for $i = 1, \ldots, m$, so that $r_{N,1} + \cdots + r_{N,m} = N$. Then, the $(N+1)$th run must be either at a new point θ_{m+1}, so m increases by 1 and the $r_{N,i}$'s for $i \leq m$ stay unchanged, or at one of the previous points θ_i, in which case only this $r_{N,i}$ increases by one and m is unchanged.

What we need now is an *infinite sequence* of points $\theta_1, \theta_2, \ldots$ in Θ such that for any finite m, the set $S_m = \{\theta_1, \ldots, \theta_m\}$ has low dispersion (relative to the size of m). Such a sequence is called a *low-dispersion sequence*. For $\Theta = [0,1]^s$, Niederreiter (1992), Theorem 6.9, gives the following low-dispersion sequence for the sup norm. Define

$$x_1 = 1, \qquad x_m = (\log_2(2m-3)) \bmod 1 \quad \text{for } m \geq 2, \tag{21}$$

where x mod 1 denotes the fractional part of x. In dimension 1, the low-dispersion sequence is defined by $\theta_i = x_i$ for all i. For $s > 1$, consider a sequence $\theta_1, \theta_2, \ldots$ such that for any integer $k \geq 1$, for $m = k^s$, the set $S_m = \{\theta_1, \ldots, \theta_m\}$ is precisely the set of all points of the form $\theta = (u_1, \ldots, u_s)$ with $u_j \in \{x_1, \ldots, x_k\}$ for $1 \leq j \leq s$. The order of the θ_i in the sequence does not matter, as long as they satisfy the above condition for all $k \geq 1$. This sequence satisfies:

$$\lim_{m\to\infty} m^{1/s} d_\infty(S_m, [0,1]^s) = \frac{1}{\ln 4}. \tag{22}$$

For $s = 1$, this is asymptotically optimal, in the sense that no other sequence can have a lower value for this limit. For $s > 1$, the smallest possible value of this limit is unknown, but it cannot be smaller than $1/2$. Therefore, one cannot achieve much better than for the above sequence, with respect to the sup norm.

The sequential version of our optimization algorithm, using a low-dispersion sequence $\theta_1, \theta_2, \ldots$, operates as follows. Define $m = m_N^*$ as a function of N in the same way as in the previous section. As N increases to $N+1$, if $m_{N+1}^* = m_N^* + 1$, perform a simulation run at the new evaluation point θ_{m+1} where $m + 1 = m_{N+1}^*$. Otherwise, i.e., if $m_{N+1}^* = m_N^* = m$, perform an additional run at the point θ_i with the smallest number of runs $r_{N,i}$ so far, for $i \leq m$ (in case of ties, break them arbitrarily).

Developments and bounds in the preceding section, under the stated Assumptions, still hold in this setting at times preceding acquisition of new evaluation points. This is because the dispersion $d_p(S_m, \Theta)$ is still $O(m^{-1/s})$, so Theorem 1 applies.

5 Stochastic Optimization with Adaptive Sample Sizes

In what we have seen so far, one performs (approximately) the same number of runs at each point of S_m. Thus the stochastic optimization is *non-adaptive*: It can be undertaken without regard to the observed values. But as the number of observations increases, from examination of the data it often becomes pretty clear which are the more promising points of S_m. Through sequential sampling (i.e., selection of the points $\theta_n \in S_m$ on the basis of preceding observations), there is hope for improvement over rates derived in the preceding section.

At promising points, one should collect more observations because it is with nearly-optimal points that sampling noise is more likely to lead to selection error. By contrast, if $J(\theta_{m,i})$ is far from J_N^*, it would take a relatively large error in $\hat{J}(\theta_{m,i})$ for $\theta_{m,i}$ to be mistaken for the optimal point. It is thus natural to adaptively concentrate the sampling effort on those more promising points. In the analysis to follow, an adaptive stochastic optimization method is offered which was motivated by these ideas and is consistent with criteria followed in earlier portions of this paper.

We give a refinement of Proposition 2 for the case where the number r of replications is not the same for all design points. Let r_i denote the number of replications at $\theta_{m,i}$, and define $\delta_i = J(\theta_{m,i}) - J_N^*$, for $i = 1, \ldots, m$. (In our adaptive scheme to follow, the δ_i's will be estimated by their sample means.)

Proposition 4 *Let ϵ be an arbitrary positive number. Under the same assumptions as in Proposition 2, one has,*

$$P[J(\hat{\theta}_N^*) - J_N^* > \epsilon] \quad \leq \quad \sum_{i=1}^{m} \exp\left[-r_i(\kappa/16)(\delta_i + \epsilon)^2\right]. \tag{23}$$

Proof. Take $\Psi = \{i : 1 \leq i \leq m \text{ and } \delta_i > \epsilon\}$. Clearly if $i \notin \Psi$ then $J(\hat{\theta}_N^*) - J_N^* < \epsilon$. Therefore, $P[J(\hat{\theta}_N^*) - J_N^* > \epsilon] = P[\theta_{m,i} \text{ is chosen, with } i \in \Psi]$. It is shown first that if $\theta_{m,i}, i \in \Psi$ is chosen, then either

$$|\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| \quad > \quad \delta_i/2 > (J(\theta_{m,i}) - J(\theta_{m,i^*}) + \epsilon)/4 \tag{24}$$

or

$$|\hat{J}(\theta_{m,i^*}) - J(\theta_{m,i^*})| \quad > \quad \delta_i/2. \tag{25}$$

If $\theta_{m,i}$ is chosen as $\hat{\theta}_N^*$, by definition (3) it must be that $\hat{J}(\theta_{m,i^*}) \geq \hat{J}(\theta_{m,i})$, which implies

$$\begin{aligned}(\hat{J}(\theta_{m,i^*}) - J(\theta_{m,i^*})) + (J(\theta_{m,i}) - \hat{J}(\theta_{m,i})) \quad &\geq \quad \hat{J}(\theta_{m,i}) + (J(\theta_{m,i}) - \hat{J}(\theta_{m,i})) - J(\theta_{m,i^*}) \\ &= \quad J(\theta_{m,i}) - J(\theta_{m,i^*}) \geq 0. \end{aligned} \tag{26}$$

By the triangle inequality, a necessary condition for (26) is that

$$|\hat{J}(\theta_{m,i^*}) - J(\theta_{m,i^*})| + |J(\theta_{m,i}) - \hat{J}(\theta_{m,i})| \quad \geq \quad J(\theta_{m,i}) - J(\theta_{m,i^*}) = \delta_i. \tag{27}$$

For this to happen, at least one of the terms on the left must be at least half as large as the right side. From the definition of Ψ, we have $\delta_i > \epsilon$, which completes showing that either (24) or (25) must happen.

Let us designate the event that (24) holds by $E(i)$, and the event that (25) holds by $E^*(i)$ regardless of i. From the preceding developments, we conclude that if the event "$J(\hat{\theta}_N^*) - J_N^* > \epsilon$" occurs then $E(i)$ or $E^*(i)$ occur for the chosen $\theta_{m,i}, i \in \Psi$. That is, if $I(i)$ is the indicator function that i is chosen,

$$P[J(\hat{\theta}_N^*) > J_N^* + \epsilon] \leq \sum_{i \in \Psi} \left(P[E(i)] + P[E^*(i)]\right) I(i)$$

Under Assumption A1 and $\delta_i > \epsilon > \epsilon/2$, $P[E^*(i)]$ is uniformly majorized by

$$P[E^*(i)] \leq P[|J(\theta_{m,i}) - \hat{J}(\theta_{m,i})| > \epsilon/4] \leq e^{-r_{i^*}\epsilon^2\kappa/16}$$

and using the elementary fact that the probability of a union of events does not exceed the sum of the probabilities of the events themselves, we have that

$$\begin{aligned} P[J(\hat{\theta}_N^*) > J_N^* + \epsilon] \quad &\leq \quad \sum_{i \in \Psi} P[E(i)] + e^{-r_{i^*}\epsilon^2\kappa/16} \\ &\leq \quad \sum_{i=1}^{m} \exp(-(J(\theta_{m,i}) - J(\theta_{m,i^*}) + \epsilon)^2 \kappa r_i/16). \end{aligned}$$

□

The setting for this section is that somehow one has selected the total number, call it N, of observations to be made in the stochastic minimization effort. We will let $n = 1, 2, \ldots, N$ indicate the current number of observations (replications) that have been collected up to the present decision time. The decision to be made is which value $\theta \in S_m$ is to be selected for the next (i.e. $(n+1)$st) observation. The basis of the adaptive stochastic minimization considered here is to choose the numbers of replications r_i adaptively, for increasing n, on the basis of previous choices of $\theta \in S_m$, so as to minimize the probability bounds given by (23). The next proposition gives the optimal replication allocation for minimizing the bound of Proposition 4, provided the numbers $\delta_i = J(\theta_{m,i}) - J(\theta_{m,i^*})$ and κ are somehow known. Following that, a strategy for inferring needed values will be offered. For economy of notation in the next development, we define,

$$K_i = (\delta_i + \epsilon)^2 \kappa / 16. \tag{28}$$

Proposition 5 *For given positive constants $K_1, \ldots, K_m$, the minimizer of*

$$\sum_{i=1}^{m} \exp[-r_i K_i] \tag{29}$$

over non-negative real vectors $(r_1, \ldots, r_m)$ subject to $\sum_{i=1}^m r_i = N$ is given by

$$r_i \;=\; \frac{\ln K_i}{K_i} + \frac{N - \sum_{j=1}^m (\ln K_j)/K_j}{K_i \sum_{j=1}^m 1/K_j}. \tag{30}$$

Proof. The relation (30) follows by writing the first order optimality conditions, using a Lagrange multiplier. For ∇ representing the gradient with respect to $(r_1, \ldots, r_m)$, we have that for some number λ,

$$\nabla \sum_{i=1}^{m} \exp[-r_i K_i] + \lambda(1, \ldots, 1) = 0.$$

The solution requires that

$$K_i \exp[-r_i K_i] = \lambda$$

for all i. Taking the logarithm and solving for r_i, one obtains

$$r_i = (\ln K_i - \ln \lambda)/K_i.$$

Combining this with the constraint $\sum_{j=1}^m r_j = N$ to eliminate $\ln \lambda$, (30) follows after easy manipulations. □

At the optimum in the previous proposition, $r_i K_i - \ln K_i$ is the same for all i. This motivates the following methodology for the case of fixed m: *At each sample time n, choose the test point $\theta_{m,i}$ for which*

$$r_i \hat{K}_i - \ln \hat{K}_i = r_i (\hat{\delta}_i + \epsilon)^2 \kappa/16 - \ln((\hat{\delta}_i + \epsilon)^2 \kappa/16) \tag{31}$$

is minimal, where $\hat{\delta}_i$ is our current sample-mean estimate of δ_i. Following this selection strategy, aside from sampling error and integer discretization, at each time the allocation will be optimal with respect to the probability bound criterion. (It will be argued later that with S_m fixed, asymptotically in n the sampling error becomes negligible.) By contrast, the non-adaptive strategy allots just as many replications to "bad" points as to promising ones. Principles for using plug-in estimators in place of the δ_i*'s* will be offered after we consider an example under idealized conditions.

Example 6 If $J(\theta)$ has at least two derivatives, then in the neighborhood of any minimum, $J(\theta)$ will resemble a quadratic. Thus the following idealized example, presuming the δ_i's are known, is heartening. (In the computational experiment afterwards, we will compare the non-adaptive strategy against the adaptive scheme with plug-in estimates of the δ_i's, for the quadratic in this example.)

Take $s = 1$, $\Theta = [0, 1]$, $J(\theta) = \theta^2$, $\epsilon = 1/m^2$, and without loss of generality, $\kappa = 16$. According to the non-adaptive (NA) strategy, ignoring sampling error, we have that for all i, $r_i = N/m$, $\theta_{m,i} = i/m$, and so our best bound is

$$\phi_{NA}(N, \epsilon) = P_{NA}[J(\hat{\theta}^*_N) - J^*_N > 2\epsilon] \;\leq\; \sum_{i=0}^{m-1} \exp(-(N/m)[(i+1)/m]^4) \;>\; \exp(-N/m^5).$$

For analytic convenience, in investigating our adaptive (A) strategy, we will ignore the logarithmic term and approximate (31) by the (suboptimal) rule: Sample next at $\theta_{m,j}$ for j the minimizer over $1 \leq i \leq m$ of

$$r_i K_i = r_i(\delta_i + \epsilon)^2. \tag{32}$$

Proposition 4 yields the bound

$$\phi_A(N, \epsilon) = P_A[J(\hat{\theta}^*_N) - J^*_N > \epsilon] \leq \sum_{i=1}^{m} \exp(-r_i K_i) \leq m \exp(-N/(m^4 Q)), \tag{33}$$

where $m^4 Q = \sum_{v=1}^{m} 1/K_v$, or $Q = 1/\sum_{v=1}^{m} v^{-4}$, which is bounded in m.

Consequently, the terms in the adaptive exponents pick up a factor of m in growth, over the nonadaptive counterparts, in these error bounds. To the extent that these upper bounds are tight, the probability of misclassification ought to be significantly smaller in the adaptive case.

6 Computational Considerations and Experiments for the Adaptive Case

We define the approximation

$$\hat{\delta}_i = \hat{J}(\theta_{m,i}) - \hat{J}(\hat{\theta}^*_n),$$

where $\hat{J}(\theta_{m,v})$ is the running sample average of observations taken up through the nth decision time at the point $\theta_{m,v} \in S_m$. Toward assessing error of approximation by using the alias $\hat{\delta}_i$ in place of δ_i in (31), suppose that for some arbitrary $\epsilon > 0$ and all $1 \leq i \leq m$,

$$|\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| < \epsilon. \tag{34}$$

Then for each i, $|\delta_i - \hat{\delta}_i| < 2\epsilon$. This is because

$$\begin{aligned} |\hat{\delta}_i - \delta_i| &= |(\hat{J}(\theta_{m,i}) - \hat{J}(\hat{\theta}^*_n)) - (J(\theta_{m,i}) - J(\theta_{m,i^*}))| & (35) \\ &\leq |\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| + |\hat{J}(\hat{\theta}^*_n) - J(\theta_{m,i^*})|. & (36) \end{aligned}$$

But under (34)

$$\hat{J}(\hat{\theta}^*_n) \geq J(\hat{\theta}^*_n) - \epsilon \geq J(\theta_{m,i^*}) - \epsilon$$

and by the choice of $\hat{\theta}^*_n$,

$$\hat{J}(\hat{\theta}^*_n) \leq \hat{J}(\theta_{m,i^*}) \leq J(\theta_{m,i^*}) + \epsilon.$$

Thus the final term in (36) is bounded by ϵ and if the event (34) holds, then indeed $|\hat{\delta}_i - \delta_i| < 2\epsilon$ for every i. The probability of the event (34) failing is majorized by

$$P[\max_i |\hat{J}(\theta_{m,i}) - J(\theta_{m,i})| \geq \epsilon] \leq \sum_{v=1}^{m} \exp\left[-r_v \kappa \epsilon^2\right],$$

thanks to Assumption (A1) and because the probability of a union of events is bounded by the sum of probabilities of the events. Since r_v would increase without bound by our selection rule if n were unbounded, the estimates are weakly consistent, in this sense.

On another computational matter, we suggest that the criterion (31) be approximated by selection of $\theta_{m,i}$ with index i the minimizer of

$$r_i(\hat{\delta}_i + \epsilon)^2 \tag{37}$$

for the next point to be observed. This rule is equivalent to ignoring the logarithmic term in the ideal rule, (31). The advantage of this latter formula is that the (usually unknown) large-deviation parameter κ cancels out. If $q \geq 2$, the approximation is justified by observing that in (31), the term $\sum_{v=1}^{m} |\ln(K_v)|/K_v \leq m|\ln(\Delta^4)|/\Delta^4$ where Δ is the dispersion. From developments in Sections 2 and 3, we have that $\Delta = \Delta(N) = O(N^{-1/(2q+s)})$ and $m = m(N) = O(N^{s/(2q+s)})$ and so $\sum_{v=1}^{m} \ln(K_v)/K_v = o(N)$ and the term $n/(K_i \sum 1/K_j)$ is dominant as n approaches N. In any case, since (37) is suboptimal, conclusions in the examples to follow are no worse than what would be expected under (31).

In summary, the data-driven approximations for implementing the adaptive stochastic optimization scheme are

1. Approximate $\delta_i = J(\theta_{m,i}) - J^*_N$ by the corresponding sample means, $\hat{J}(\theta_{m,i}) - \hat{J}^*_n$ where we have defined $\hat{J}^*_n = \min_{1 \leq v \leq m} \hat{J}(\theta_{m,v}) = \hat{J}(\hat{\theta}^*_n)$.

2. Use the convenient criterion: $\arg\min_i r_i(\hat{\delta}_i + \epsilon)^2$ in place of (31), thereby avoiding the task of guessing or bounding κ.

Hopefully, the computational examples offered below are somewhat representative of the performance that can be anticipated.

Example 7 Some comparative simulation experiments are reported here. We test the two target functions $J(\theta) = \theta^2$, as in Example 6 and $J(\theta) = \theta \sin(50\theta)$, with results reported in Tables 2 and 3, respectively. In these computations, $\Theta = [0, 1]$.

Each observation is additively corrupted by a $N(0, 1)$ observation. For the adaptive rule, the threshold ϵ in (31) is taken to be the tolerance $(N/\ln(N))^{-2/5}$, in accordance with Theorem 1. The program begins by making one observation at each point in S_m, and thereafter follows the adaptive strategy.

Each entry in these tables is based on 10 replications. The entries labelled "Ave" are the sample averages (over the 10 replications) of the of selection errors $J(\hat{\theta}_N^*) - J_N^*$ at the declared best design points, and following that are the discretization errors $J_N^* - J^*$. The column labelled $\hat{\sigma}(J(\hat{\theta}_N^*))$ provides the sample standard deviations. We have also included the values of the allowed error threshold as in Theorem 1, equation (17), for the values N. We examined the actual errors in the replication blocks and the column labelled "# Bad" records the number of replications in which this threshold was exceeded. In the listing, in the Tables, "A" designates Adaptive and "NA" stands for Non-Adaptive. In our simulations, we began by sampling at each point in S_m once, and thereafter reverting to the adaptive search.

In scanning the results of these experiments, there is clear evidence that with increasing sample size, adaptation is reducing the number of exceedances. Since the scaling of the error thresholds is arbitrary, these numbers are only suggestive. What is more suggestive is that (aside from the $N = 100$ case) the sample averages of the selection errors are smaller for the adaptive rule, which would indicate that whatever scaling of tolerance is used, adaptation has the advantage. Even in the case of the highly oscillatory amplified sine function, it is selection rather than discretization error that dominates, for larger N, and thus improvement through adaptive sample sizes results in increased accuracy of the declared optimizer.

Example 8 Here we return attention to the two-dimensional computation comprising Example 5. The simulation and setup is as described there. The difference here is that the adaptive stochastic minimizer is used. Approximately the same range of N is employed. The number of $m = m(N)$ of test points is taken to be approximately $10N^{2/5}$, in accordance with the dispersion theory (the approximation being that m must be a perfect square). In the table, a choice $\hat{\theta}_N^*$ is deemed "bad" if it represents an exceedance of $(N/\ln N)^{-2/5}$. This is more stringent than the tolerance used in the non-adaptive example, but from extension of ideas in Example 6, convergence can be assured in the quadratic case.

In comparing Table 4 with the corresponding summary of the non-adaptive case (Table 1) we see that in this case, adaptation leads to markedly-improved performance. The number of exceedances at larger N has fallen, despite the reduced tolerance and despite more grid points being used at each level. The increase in grid-point density has led to lowering the discretization error $J_N^* - J^*$. At the higher levels of N, the average selection error $J(\hat{\theta}_N^*) - J_N^*$ has fallen way below the corresponding entries of Table 1.

Table 1: Computational Illustration of the Stochastic Optimizer: Nonadaptive Case

Observation: $L(\theta) = (\theta_1 - 0.5) * \sin(10 * \theta_1) + (\theta_2 + 0.5) * \cos(5 * \theta_2) + N(0,1)$

N	# Bad	$\mathrm{Ave}(J(\hat{\theta}_N^*) - J_N^*)$	$J_N^* - J^*$	$\hat{\sigma}(J(\hat{\theta}_N^*))$	$(N/\ln N)^{-1/3}$	m
75	2	0.2510	0.1402	0.2282	0.3584	25
396	3	0.1269	0.0249	0.1658	0.2316	36
931	2	0.1230	0.0446	0.1090	0.1904	49
4719	3	0.0649	0.0082	0.0993	0.1194	81
9700	4	0.0560	0.0417	0.0439	0.0973	100
19118	1	0.0414	0.0438	0.0549	0.0791	121

Table 2: Comparison of the Performances of the Non-Adaptive and Adaptive Stochastic Optimizers: I. Quadratic Case.

Observations: $L(\theta) = \theta^2 + N(0,1)$

N	Method	# Bad	$\mathrm{Ave}(J(\hat{\theta}_N^*) - J_N^*)$	$J_N^* - J^*$	$\hat{\sigma}(J(\hat{\theta}_N^*))$	$(N/\ln N)^{-2/5}$	m
100	NA	1	0.069	0.0100	0.1111	0.2919	10
100	A	2	0.142	0.0100	0.2030	0.2919	10
500	NA	2	0.079	0.0021	0.0882	0.1729	22
500	A	0	0.079	0.0021	0.0681	0.1729	22
1000	NA	0	0.057	0.0010	0.0514	0.1367	31
1000	A	0	0.021	0.0010	0.0301	0.1367	31
5000	NA	0	0.059	0.0002	0.0214	0.0781	70
5000	A	0	0.014	0.0002	0.0110	0.0781	70

Table 3: Comparison of the Performances of the Non-Adaptive and Adaptive Stochastic Optimizers: II. Amplified Sine Function Case.

Observations: $L(\theta) = \theta \sin(50\theta) + N(0,1)$

N	Method	# Bad	$\mathrm{Ave}(J(\hat{\theta}_N^*) - J_N^*)$	$J_N^* - J^*$	$\hat{\sigma}(J(\hat{\theta}_N^*))$	$(N/\ln N)^{-2/5}$	m
100	NA	6	0.250	0.3812	0.226	0.2919	10
100	A	3	0.130	0.3812	0.219	0.2919	10
500	NA	5	0.169	0.2668	0.179	0.1729	22
500	A	1	0.037	0.2668	0.115	0.1729	22
1000	NA	3	0.039	0.0517	0.062	0.1367	31
1000	A	1	0.013	0.0517	0.041	0.1367	31
5000	NA	0	0.045	0.0010	0.075	0.0781	70
5000	A	0	0.014	0.0010	0.011	0.0781	70

Table 4: Computational Illustration of the Adaptive Stochastic Optimizer: Adaptive case

Observations: $L(\theta) = (\theta_1 - 0.5)\ \sin(10\ \theta_1) + (\theta_2 + 0.5)\ \cos(5\ \theta_2) + N(0,1)$

N	# Bad	$\text{Ave}(J(\hat{\theta}_N^*) - J_N^*)$	$J_N^* - J^*$	$\hat{\sigma}(J(\hat{\theta}_N^*))$	$(N/\ln N)^{-2/5}$	m
100	3	0.1677	0.0446	0.1884	0.2919	49
500	1	0.0828	0.0417	0.0862	0.1729	100
1000	1	0.0607	0.0249	0.0520	0.1367	144
5000	0	0.0113	0.0070	0.0148	0.0781	289
10000	1	0.0137	0.0208	0.0277	0.0611	361
20000	0	0.0055	0.0064	0.0114	0.0476	484

7 Conclusions

The objective of the present paper is to explore the influence of smoothness assumptions in the context of stochastic minimization. Notions of dispersion and deviation theory have served these ends; the findings have included prescriptions for the numbers and (non-adaptive) locations of sampling points within a continuous risk-function domain in a Euclidean space. As noted, these findings impinge on developments in the area of ranking and selection; thus our techniques can deal with minimization of a noisy function over a continuum of values, and our results do not depend on normality assumptions. Since the concepts are based on "model-free" or "machine-learning" approaches, the methodology here is appropriate for on-line experimentation and optimization as well as stochastic optimization through simulation.

Both fixed and adaptive sampling strategies have been considered. The latter case impinges on the literature of nonparametric bandit theory and stochastic approximation. Thus in contrast to the latter discipline, we can assure global convergence without unimodality assumptions. To our knowledge, bandit theory has not included topological assumptions, as we have here, and as a consequence, rates and bounds established in the present work are superior to developments in the literature. Theory and experimentation clearly show that when feasible, adaptive selection gives improved performance.

Remaining issues include extension to steady-state models and to examination and weakening of assumptions, particularly those used for the adaptive selection criterion. Finally, for the case as in Section 4, of dynamically increasing the set S_m, there is the intriguing issue of how to place new points θ adaptively, as evidence of promising decision regions accumulates. In learning theory, a central and largely unresolved problem is how to use past values to select promising regions. The adaptive methodology suggests a novel way to automate this task: One constructs a dense grid dynamically as in Section 4 over the entire search region. The $r_i(\hat{\delta}_i + \epsilon)^2$ criterion should, theoretically, automatically allot replications where values seem promising and ignore other domains. More investigation of this insight will be undertaken.

Acknowledgments

This work has been supported by NSERC-Canada grants No. ODGP0110050, SMF0169893, and WFA0139015, and by FCAR-Québec grant No. 93ER1654 to the second and third authors. The comments of the Area Editor Barry Nelson and of two anonymous referees helped us to improve the presentation greatly.

References

Bechhofer, R., T. Santner, and D. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* New York: Wiley.

Benveniste, A., M. Métivier, and P. Priouret. 1990. *Adaptive Algorithms and Stochastic Approximation.* New York: Springer-Verlag.

Conway, J. H., and N. J. A. Sloane. 1988. *Sphere Packings, Lattices and Groups.* Grundlehren der Mathematischen Wissenschaften 290, New York: Springer-Verlag.

Deheuvels, P. 1983. Strong bounds for multidimensional spacings. *Z. fur Wahrsch. Verw. Geb.*, **64**, 411–424.

Dippon, J. and V. Fabian. 1994. Stochastic approximation of global minimum points. *Journal of Statistical Planning and Inference*, **41**, 327–347.

Ellis, R. S. 1985. *Entropy, Large Deviations, and Statistical Mechanics.* Springer Verlag.

Feller, W. 1966. *An Introduction to Probability Theory and Its Applications, Vol. 2.* first ed. New York: Wiley.

Glasserman, P. 1991. *Gradient Estimation Via Perturbation Analysis.* Norwell, MA: Kluwer Academic.

Glynn, P. W. 1990. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, **33**(10), 75–84.

Härdle, W. 1989. *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Ho, Y.-C., and M. E. Larson. 1995. Ordinal optimization approach to rare event probability problems. *Discrete Event Dynamic Systems: Theory and Applications*, **5**, 281–301.

Kushner, H. J., and D. S. Clark. 1978. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Volume 26 of *Applied Mathematical Sciences.* Springer-Verlag.

Kushner, H. J., and F. J. Vázquez-Abad. 1996. Stochastic approximation algorithms for systems over an infinite horizon. *SIAM Journal on Control and Optimization*, **34**, 712–756.

Kushner, H. J., and G. Yin. 1997. *Stochastic Approximation Algorithms and Applications.* New York: Springer Verlag.

Lai, T. L., and S. Yakowitz. 1995. Machine learning and nonparametric bandit theory. *IEEE Transactions on Automatic Control*, **40**, 1199–1210.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis.* Second ed. New York: McGraw-Hill.

L'Ecuyer, P. 1991. An overview of derivative estimation. In *Proceedings of the 1991 Winter Simulation Conference*, 207–217. IEEE Press.

L'Ecuyer, P. 1993. Two approaches for estimating the gradient in a functional form. In *Proceedings of the 1993 Winter Simulation Conference*, 338–346. IEEE Press.

L'Ecuyer, P., and G. Perron. 1994. On the convergence rates of IPA and FDC derivative estimators. *Operations Research*, **42**(4), 643–656.

L'Ecuyer, P., and F. Vázquez-Abad. 1997. Functional estimation with respect to a threshold parameter via dynamic split-and-merge. *Discrete Event Dynamic Systems: Theory and Applications*, **7**(1), 69–92.

L'Ecuyer, P., and G. Yin. 1998. Budget-dependent convergence rate for stochastic approximation. *SIAM Journal on Optimization*, **8**(1). To appear.

Matejcik, F. J., and B. L. Nelson. 1995. Two-stage multiple comparisons with the best for computer simulation. *Operations Research*, **43**, 633–640.

Mukhopadhyay, N., and T. K. S. Solansky. 1994. *Multistage Selection and Ranking Procedures.* New York: Marcel Dekker.

Müller, H.-G. 1985. Kernel estimators of zeros and location and size of extrema of regression funcitons. *Scandinavian Journal of Statistics*, **12**, 221–232.

Müller, H.-G. 1989. Adaptive nonparametric peak estimation. *The Annals of Statistics*, **17**, 1053–1069.

Niederreiter, H. 1992. *Random Number Generation and Quasi-Monte Carlo Methods.* volume 63 of *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics.* Philadelphia: SIAM.

Petrov, V. V. 1975. *Sums of Independent Random Variables.* New York: Springer Verlag.

Prakasa Rao, B. L. S. 1983. *Nonparametric Functional Estimation.* New York: Academic Press.

Robbins, H. 1952. Some aspects of the design of stochastic experiments. *American Mathematical Society Bulletin*, **58**, 527–535.

Rubinstein, R. Y., and A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* New York: Wiley.

Shwartz, A., and A. Weiss. 1995. *Large Deviations for Performance Analysis.* London: Chapman and Hall.

Spall, J. C. 1992. Multivariate stochastic approximation using simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control*, **AC-37**, 331–341.

Sukharev, A. G. 1971. Optimal strategies of the search for an extremum. *Zh. Vychisl. Mat. i Mat. Fiz.*, **1**, 910–924. In Russian.

Yakowitz, S. 1993. A globally convergent stochastic approximation. *SIAM J. on Control and Optimization*, **31**, 30–40.

Yakowitz, S., and J. Mai. 1995. Methods and theory for off-line machine learning. *IEEE Transactions on Automatic Control*, **40**(1), 161–165.